УДК 81-139

Филимонов В. В., *Амиева А. М., Сергеев А. П.*
УрФУ, г. Екатеринбург, Россия

Кластеризация русскоязычных текстов с применением статистики χ^2

Аннотация

Проблема обнаружения скрытых структур текста связана с перспективной методикой установления авторства. В работе описан Корпус текстов русского языка, созданный для исследований текстов методами математической статистики. Описывается исследование, проведённое на базе Корпуса с применением статистики χ^2. Кластеризация текстов, обнаруженная в результате исследования, может служить основанием для их атрибуции. Работа выполнена на кафедре полиграфии и веб-дизайна ИРИТ-РтФ УрФУ.

Ключевые слова: корпус, частота, вероятность, статистика χ^2, функция распределения.

Filimonov V. V., *Amieva A. M., Sergeev A. P.*
UrFU, Ekaterinburg, Russia

Clustering of Russian texts using χ^2 statistics

Abstract

The problem of detection of the hidden structures of the text is associated with a promising method of attribution. The paper describes a corpus of Russian language created for investigations of texts by the methods of mathematical statistics. The research conducted on the basis of the statistics $\chi^2$is described. Clustering of texts discovered by the study can serve as a basis for their attribution. The work was coudveted at the department of printing art and web-design, Ural Federal University.

Keywords: case, frequency, probability, χ^2 statistics, distribution function.

Введение

Исследования текста связаны или с измерением его пространственных характеристик (длина строки, интерлиньяж, размер шрифта), либо с изучением смысловых единиц (предложения, фразы, и т. п.). Но они не включают в себя скрытые структурные паттерны, которые не связаны со смыслом текста или с рисунком шрифта [1].

Структурным исследованиям языка и текста посвящены работы в рамках школы структурализма.

Пражские лингвисты В. Матезиус, Й. Вахек, В. Скаличка [12] и др. устанавливали функциональную дифференциацию традиционного литературного языка. Датские структуралисты Л. Ельмслев, В. Брёндаль хотели создать универсальную и общую для всех теорию в лингвистике.

Американские учёные Э. Сепир, Л. Блумфилд и др. вырабатывали объективные процедуры формального внутреннего анализа языка при стремлении к научному обоснованию проблем комбинаторики.

Британские структуралисты М. Кёрквуд Халлидей, Ф. Р. Палмер [10] и др. работали над решением проблем совместимости идеи создания всеобщей языковой теории. По мнению французских структуралистов (А. Мартине, Г. Гийом [5], Л. Теньер и др.), язык – это целое, которое организованно и связанно.

Отечественные исследования в области структуры языка проводились в рамках московско-тартуской школы (Ю. М. Лотман, З. Г. Минц [9], В. Н. Топоров [13] и др.). Её представители проявляли интерес к проблемам структуры и принципов функционирования систем знаков в сообществе носителей языка, а также общими для всех них структурно-семиотическими методами анализа культурных текстов. В исследованиях применяли математическое моделирование, элементы теории информации и статистические методы.

Методологической основой исследований в рамках московско-тартуской школы является тезис де Соссюра о единстве структур всех существующих языков. Сам текст понимается

как последовательность знаков, которая может быть проанализирована с позиций семиотики.

Характеристики пространственной организации текстов и их влияния на чтение и понимание текстовой информации изучаются в рамках исследования удобочитаемости (Weber A., Cohn H., Артемов В. А. [2], Ушакова М. Н. [14], Гешев М. [4] и др.). Исследуется не столько структура текста как организованного целого (Gestalt), сколько удобство чтения. В контексте этих исследований его соотносят с понятием юзабилити, которое определяется в системе стандартов ISO как эффективность, результативность, удовлетворённость» [1].

В данной работе исследуется гипотеза о том, что в текстах есть скрытые структуры. Они не связаны с семантикой, синтаксисом, но, возможно, связаны с прагматикой текста. Семантическая сторона не рассматривается в силу её конвенциальности, т. к. читатель текста является его интерпретатором и соавтором. Для этого мы исследовали корпус текстов. Предметом исследования являются комбинации букв по три. При исследовании всех 33 букв русского языка, количество возможных комбинаций было бы большим. Тогда для исследований необходимы бы были тексты с длиной более сотни тысяч букв. Поэтому рассматривались только гласные буквы.

Корпус текстов русского языка

В 2015 году на кафедре «Полиграфии и веб-дизайна» ИРИТ–РтФ УрФУ силами преподавателей и студентов начато создание своего Корпуса текстов русского языка. Он предназначен для исследований статистических закономерностей в русском языке. На сегодняшний день корпус состоит из 686 текстов художественного, научного, социально-политического, административного направлений, а также из газетных и журнальных публикаций. Подкорпус художественных текстов состоит из литературных произведений XIX–XX вв. русских, советских и зарубежных авторов в русском переводе. Для переводных текстов указывается двойное авторство: автор первоначального текста и автор перевода. То же касается и научных текстов, в подкорпус которых вошли

труды крупнейших учёных XIX–XX вв., монографии, статьи. В подкорпус социально-политических текстов входят речи политических деятелей на съездах, форумах, программные статьи и т. п. Административный подкорпус составляют кодексы, законы и прочие нормативные документы. Отдельно можно выделить подкорпус текстов религиозного направления. В него входят «канонические» тексты различных конфессий, произведения авторитетных религиозных деятелей. Все тексты этого подкорпуса представлены в современных русских переводах.

Корпус текстов русского языка может быть использован в исследованиях статистических закономерностей отдельно взятого текста. Для таких исследований необходима достаточная длина каждого текста, поэтому небольшие тексты (отдельные стихотворения, газетные и журнальные заметки и т. п.) объединяются в единый комплекс. Он может состоять из нескольких десятков отдельных произведений, объединённых авторством, временем написания и темой. Стихотворения представлены сборниками. Предполагается, что сборник составлен самим поэтом, который собирал их по каким-то признакам. Если сборник компоновал редактор, то он ориентировался на тематику, периоды жизни или творчества поэта и по каким-то другим признакам.

Тексты представлены в формате *.txt. Сохранена авторская орфография, не учитывается буква «ё». Общий объём корпуса на сегодняшний день составляет 754, 552 Мб.

Исследование

В ходе исследования были проанализированы все тексты Корпуса текстов русского языка. Исследования были проведены на специально написанных программах «Coder» и «QLines».

Сначала с помощью программы «Coder» был осуществлён перевод букв в цифровой код. Кодировались только гласные буквы. Согласные буквы, пробелы и знаки препинания не вошли в кодированный текст. Буква «ё» в большинстве текстов представлена как «е», поэтому при подсчётах она так и учиты-

валась. Для 10 гласных русского языка была задана соответствующая комбинация из двух цифр. Соответствие приведено в табл. 1.

Таблица 1

Соответствие кода и буквы

Буква	Код
а	00
е	01
ё	01
и	03
о	04
у	05
ы	06
э	07
ю	08
я	09

Десять гласных букв кодируются девятью символами. После кодирования текст приобрёл следующий вид (рис. 1).

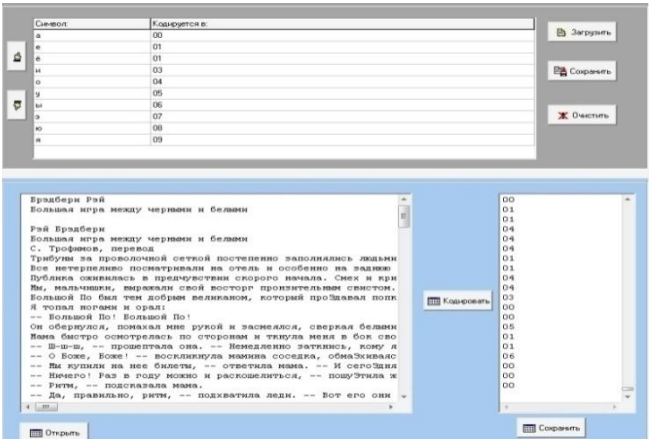

Рис. 1. Вид текста («Большая игра между чёрным и белым» Р. Брэдбери) после кодирования

Для каждого текста была подсчитана частота появления (вероятность появления) каждой гласной буквы. Частота появления отдельной буквы (p) выражается отношением количества её появлений в тексте (v) к его длине (N). Под длиной текста следует понимать общее количество гласных букв в тексте

$$p = \frac{v}{N} \qquad (1).$$

На рис. 2 представлена вероятность появления гласных в Корпусе.

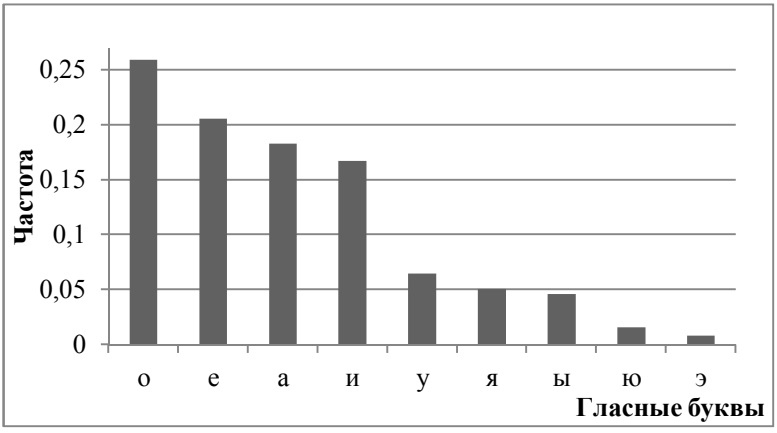

Рис. 2. Частота появления гласных букв
в Корпусе текстов русского языка

Упорядочили буквы по убыванию вероятности их появления. Полученный порядок не меняется от текста к тексту, несмотря на расхождения в числовых значениях вероятностей. Это даёт нам право считать такой порядок «частотной константой» языка. Средние вероятности появления гласных в текстах представлены в табл. 2.

Таблица 2

Средние вероятности появления отдельных гласных букв в тексте

Буква	Частота
а	0,183
е	0,206
и	0,167
о	0,259
у	0,064
ы	0,046
э	0,008
ю	0,016
я	0,051

На следующем этапе исследовались вероятности появления комбинаций букв по три буквы. Вероятность появления комбинации – это произведение вероятностей появления букв, входящих в неё. Такое представление имеет смысл, если предположить, что буквы в тексте появляются независимо друг от друга. Полученное таким способом значение вероятности назовём «теоретической вероятностью».

$$p_i = p_{i_1} \cdot p_{i_2} \cdot p_{i_3} \qquad (2),$$

где p_i – вероятность появления комбинации, p_{i_1}, p_{i_2}, p_{i_3}– вероятности появления первой, второй, третьей букв в комбинации.

Возможное количество вариантов троек: $9^3 = 729$ (число комбинаций из девяти гласных, взятых по три).

Далее при помощи «QLines» были пересчитаны все тройки гласных во всех текстах Корпуса. Полученные значения будем называть «экспериментальным количеством».

Теоретическое и экспериментальное количество троек гласных в тексте отличаются друг от друга. Произведение вероятностей появления отдельных гласных не может быть равно нулю. Но в реальном тексте могут отсутствовать некоторые возможные варианты троек. Как правило, в тексте отсутствуют около ста из семисот двадцати девяти возможных комбинаций.

Для оценки отличия предполагаемого от действительного распределения применяется статистика χ^2.

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(p_i^{\text{theor}} - p_i^{\text{emp}}\right)^2}{p_i^{\text{theor}}} \qquad (3),$$

где p_i^{theor} вычисляется по формуле (2), а p_i^{emp} было получено из реального текста в программе «QLines».

При использовании количественных значений при вычислении критерия Пирсона, его величина зависит от количества исследуемых комбинаций, то есть χ^2.

Количество гласных умножали на 50000 и это произведение делили на действительную длину текста.

По полученным данным была построена накопительная функция. График представлен на рис. 3.

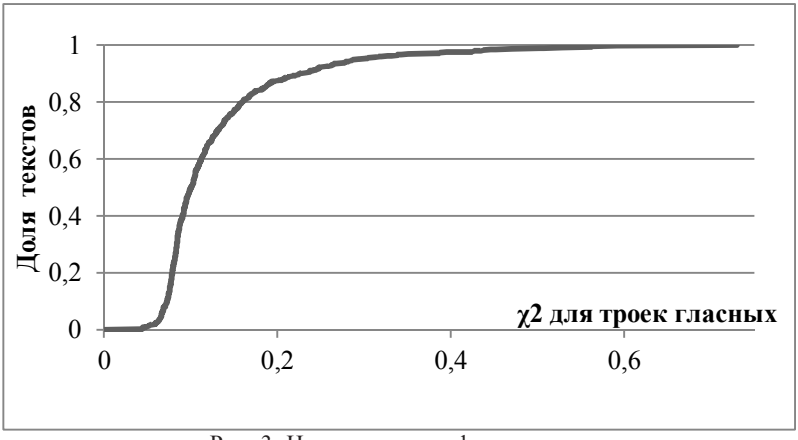

Рис. 3. Накопительная функция

Используя метод разложения функции в ряд Тейлора, продифференцировали накопительную функцию (формулы 4 и 5) и построили плотность вероятности для критерия Пирсона (рис. 4):

$$f'(x_1) = \frac{f(x_1) - f(x_2)}{x_2 - x_1} \qquad (4),$$

$$f'(x_2) = \frac{f(x_1) - 3 \cdot f(x_2) + 3 \cdot f(x_4) - f(x_5)}{3 \cdot (x_4 - x_1)} \qquad (5),$$

где значения переменной x – значения χ^2, полученные с помощью программы «QLines», а значения $f(x_n)$ –доля текстов, посчитанная для построения накопительной функции (рис. 3).

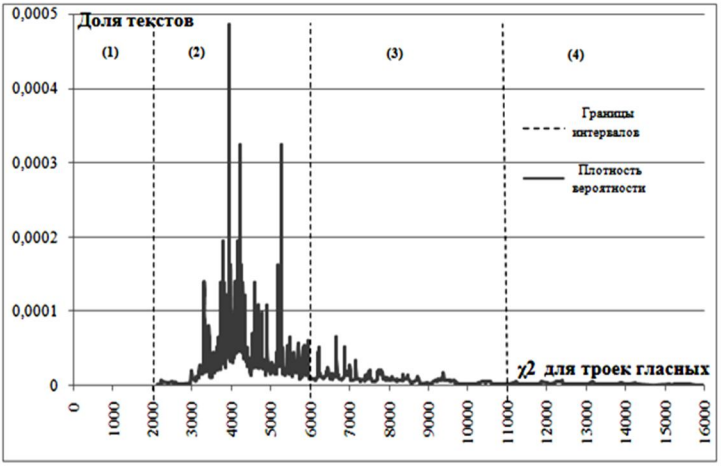

Рис. 4. Плотность вероятности для критерия Пирсона

Выводы

График плотности вероятности представляет собой линейчатый спектр. Рассмотрели, каким образом на нём расположились тексты. Для этого сопоставляли значения χ^2 с текстом, который ему соответствует. Выяснилось, что тексты расположились определённым образом.

В интервале (1) от 0 до 2000 нет ни одного текста. В интервале (2) от 2000 до 6000 встречаются только художественные тексты. В нём наблюдается и большинство ярко выраженных пиков. Эти пики соответствуют романам и повестям.

В интервале (3) от 6000 до 11000 кроме художественных появляются научные тексты, монографии и статьи, а также труды В. И Ленина, К. Маркса, тексты выступлений

Н. С. Хрущёва, М. С. Горбачёва.

В последнем интервале (4) более 11000 встречаются административные тексты: Конституция РФ, различные кодексы и законы.

Художественные тексты составляют большинство и встречаются в трёх интервалах.

Таким образом, можно говорить о том, что тексты Корпуса объединяются в несколько кластеров. Эта кластеризация связана с величиной χ^2, т. е. можно предположить, что в тексте есть некоторые скрытые структурные элементы, которые не связаны со смыслом текста. Предполагается, что дальнейшие исследования в этом направлении позволят представить скрытые структуры текста и их элементы в явном виде [1].

Список литературы

1. Амиева А. М., Филимонов В. В., Сергеев А. П. Основные методики исследования структуры текста // Передача, обработка, восприятие текстовой и графической информации. Екатеринбург, 2015. С. 251–263.

2. Артёмов В. А. Технографический анализ суммарных букв нового алфавита // Письменность и революция. MJL, 1933. № 1. С. 58–76.

3. Гешев М. Я., Колосов А. И. Влияние некоторых технологических факторов набора на удобочитаемость текстов // Научные труды по технологии полиграфического производства. М. : Моск. полигр. ин-т, 1973. № 20. С. 18–21.

4. Гийом Г. Принципы теоретической лингвистики. М.: Прогресс, 1992.

5. Корпусная лингвистика [Электронный ресурс]. Режим доступа: http://corpora. iling. spb. ru (дата обращения 23.11.2015).

6. Минц З. Г. Структура предложения и типология художественных текстов // Летняя школа по вторичным моделирующим системам. Кяэрику, 10-20 мая 1968 г.: Тезисы. Тарту, 1968. С. 93–100.

7. Палмер Ф. Р. Настроение и модальность. Издательство Кембриджского университета, Кембридж (1986), ISBN 0-521-31930-7.

8. Скаличка В. Асимметрический дуализм языковых единиц // Пражский лингвистический кружок. М., 1967.

9. Топоров В. Н. Пространство и текст // Текст: семантика и структура. М., 1983. С. 227–284.

10. Ушакова М. Н. Новый шрифт для газет // Полиграфическое производство. 1952. № 4. С. 22–23

11. Филимонов В. В., Живодеров А. А., Горбич Л. Г. Экспрессия и упорядоченность в письменной речи // Известия УрФУ. Серия 1. Проблемы образования, науки и культуры. 2012. № 3(104). С. 313–319.